Gray Areas of Assessment Systems

NCEO Synthesis Report 32

Published by the National Center on Educational Outcomes

Prepared by:

Patricia Almond
Oregon Department of Education

Rachel Quenemoen
National Center on Educational Outcomes

Kenneth Olsen
Mid-South Regional Resource Center

Martha Thurlow
National Center on Educational Outcomes

March 2000

Executive Summary

As part of our nation's educational commitment to equity and excellence for all, we must develop better understanding of what it means to be accountable for all children, and identify more inclusive strategies of assessment and accountability. In response to our national commitment, and to specific legislation such as Title I of the Improving America's Schools Act (IASA) and the Individuals with Disabilities Education Act 1997 (IDEA '97), states and school districts are in the

midst of developing large-scale assessment systems. Some have considered the challenge of students who do not fit into these assessment systems as one of "gray area students." New understanding is emerging that the problem does not lie with the students, but with the systems.

This paper clarifies what is meant by "gray areas of assessment" systems, delineates the primary issues that surround and contribute to gray areas, and provides suggestions for developing fully inclusive systems. We provide brief case studies of the assessment practices in two states, thereby highlighting the reality of gray areas as states implement their assessment systems. After a review of the national reform context, we present a model that provides a basis for defining and addressing gray area concerns.

Five interrelated questions are posed to define and address gray area concerns in any state or district at any point in time:

- What is driving large-scale assessment programs, and how does that affect gray area concerns?
- How does a state or district approach to content and performance standards affect gray area concerns?
- How do test accommodation and modification policies affect gray area concerns?
- To what extent do assessment formats affect gray area concerns?
- How does the nature of the high and low stakes accountability system affect gray area concerns?

We explore each of these questions by first identifying the context of each question, and then identifying issues to address and discuss.

As more states address these issues, it will become clear that the gray areas are not the same everywhere. The number of issues and nature of those issues are related to the state or district context, and therefore will be different in different places. Only by beginning this identification of relevant issues and responding to them can states and districts hope to avoid the criticism that their assessment systems do not account for every student within their public education system.

Many current assessment systems do not account for every student within our public school system. As a result, our nation’s understanding about how all students are achieving and how all schools are doing may be distorted and incomplete.

Overview

As a nation, we are committed to a goal of all students learning to high standards, and we have developed assessment systems to measure our progress. But an alarming critique of these assessment systems has emerged from several groups. This critique states, "Many current assessment systems do *not* account for every student in our public school system. As a result, our understanding about how all students are achieving and how all schools are doing may be

distorted and incomplete."

Many state assessment directors may cringe at this criticism, and a few will even deny the allegation. Some commercial test publishers may denounce the critique, explaining that it holds little relevance to the systems that test publishers produce and distribute. Even a number of researchers and psychometricians may warn us that tests should not be asked to do things that they were not designed to do; this, in a way, is a justification of why most large-scale assessments are not appropriate for all students.

Despite these objections, the critique of current assessment practices cannot be summarily dismissed. We rely on these assessment systems to report how schools are doing in educating all students. Discomfort with or denial of the critique should not stop us from asking some tough questions and exploring answers that may differ depending on what beliefs and perspectives are brought to the issue. Resorting to the contention that there are students who just do not fit into the assessment system, as though there was something wrong with the students, is not the solution either.

We think the situation is reminiscent of another assessment scenario that has gained national attention—the Lake Wobegon effect. Although we chuckle when Garrison Keillor describes Lake Wobegon, "where all of the children are above average," we know that statements like his can be true only if we do not include "all of the children" in the assessment results. Those who are not above average just do not count; they are not even considered.

Is it true that there are a number of public school students who do not take accountability tests? Yes. We now know that there are students in every public school who traditionally have been exempted from large-scale assessments for a variety of reasons. They may have disabilities, be English language learners (ELL), or may be in alternative placements. In addition, there are students who may take the tests, but whose scores either are not counted or do not adequately reflect their performance. For example, there are a number of students who take the tests, but who cannot respond adequately to them. Their instructional level may be far below the level of the test, or the accommodations they are allowed to use are not the ones that they need to really show what they know.

Nonetheless, we do have data that provide a picture of how students in our schools are performing. We know that our nation has been intensely scrutinizing academic achievement since *A Nation at Risk* first appeared in 1983. Ever since that historic report, data have been amassed and used to report trends on the academic performance of public school students (e.g., from the National Assessment of Educational Progress [NAEP], and from international assessments like the Third International Mathematics and Science Study [TIMSS]). These historical data on schooling provide useful barometers in tracking trends in education, and in framing needed reforms. Thus, from this viewpoint, we have good data on how students are achieving and how schools are doing. If we start from the belief that we have a good set of data, it is very difficult to accept a statement that those data do not account for every student.

Even with federal legislation designed to push accountability for all (e.g., the Improving America's Schools Act and the Individual with Disabilities Education Act amendments of 1997) we may be unable to reverse the current practice of partial accountability unless we can build a bridge from the past to the present, and to the future. If we want to advance our understanding of learning in the face of our nation's commitment to equity and excellence for all, we must figure out what it means to be accountable for all children, and then explore more inclusive strategies. Those students typically left out of the total picture have been referred to by some as the "gray

area students." More recently, this term has been redefined as an issue of the "gray areas of assessment" (NCEO, 1999).

The purpose of this paper is to clarify what is meant by "gray areas of assessment" systems, to delineate the primary issues that surround and contribute to gray areas, and to provide suggestions for developing fully inclusive systems. We hope that test publishers, state testing directors, state special education personnel, and researchers will use this discussion to launch a productive dialogue. We hope to trigger solutions that will help us move to truly inclusive accountability systems.

Examples of How Gray Area Assessment Issues Affect State Implementation

At the state level, gray area issues play out in complex and interrelated ways. We have developed two illustrative case studies, based on experiences in several states. The case studies are composite studies, and do not correspond precisely to any single state.

State A

In one state the adopted standards were clearly established at the 3rd, 5th, 8th, and 10th grades. The test was developed narrowly, to measure student performance in relation to these standards. To do so accurately the state had developed an assessment with a majority of test items close to the standard. The one to two per cent of students working on a life skills curriculum would be taking a test that "fit" their goals and objectives, their instruction. Students in the 3rd grade would take the third grade test and learn whether they "met" the benchmark standards. Students in the 5th grade would do the same, and so on. The problem that surfaced was that it appeared there were students who would not be able to take either test. On the one hand the students were not close enough to the standard to take the benchmark test, and on the other hand, a life skills assessment was not appropriate for them. These students fell in between the two tests. People began using the term "gray" to describe the students because the students fell in a never-never land where neither assessment would be appropriate.

For example, in reading, the state benchmark assessment focuses on reading a passage and typically answering four or five comprehension questions about the passage. At third grade some low performing students are still working on decoding. Beginning reading skills prevent them from reading the passage. They also have difficulty reading the questions, and even more difficulty identifying the correct answer. What these students need is an easier reading test. In their state, there has been no easier reading test. In fact, early in the reform there was a general objection to developing an easier reading test. There was a fear that it would send the wrong message and be seen as an attempt to lower the newly adopted high performance standards. With the reading test the state can successfully tell each student whether they "met" the standard. Students who cannot take the test do not "meet" the standard. Instructionally, the students are working on the right material. It is the assessment that does not "fit" the student. In the gray area between the large-scale assessment and the alternate assessment, there was no test to measure how well a student was doing on the standards, no test to tell students how far they still had to go to meet the standard.

In this state, it became clear that gaps existed between the "ideal" assessment system, and the emerging reality. These are portrayed in Figure 1.

Figure 1. Gaps between "Ideal" Assessment System and the Emerging Reality for State A

	An Emerging Reality in State A
	Students who take the general education test with comparable scores
	Students who take the general education test accommodated with comparable scores
Gray Area	Students who take the general education test modified without comparable scores
	Students who do not take any test because the alternate assessment does not address their curriculum and the general education assessment even with modifications is too difficult
	Students who take the alternate assessment

From this approach the state concluded that the gray area is the area where kids do not count in the accountability system, and where valid data do not exist on which to base school improvement plans. Either there is no assessment available at the students' levels of performance, or scores from the assessment do not count in the accountability reporting system.

To be both comprehensive and inclusive, the assessment system needs to change. (Note: For this example state, "comprehensive" means provides information about achievement on the standards for each and every student in the population; "inclusive" is the opposite of exclusive and indicates that the all students without exception are able to and are expected to participate in the test.)

But how should it change? One possibility is that the state could develop scoring and reporting methods for modified tests that would render the scores comparable for the system's stated purposes. If this could be done, one gray area of the assessment system would be removed. The other area involves content and performance levels that are not currently addressed by the system. Here there are several possibilities:

- Allow students to take lower level assessments;
- Widen the range of the tests by adding items further from the cut scores, then more students would be able to take them;
- Revise student instructional programs to more aggressively help students reach the range that the test covers;
- Re-examine modifications to see whether students would be able to take the test with allowable accommodations after all; or
- Increase the allowable modes of responding for students who cannot take the

test even with modifications.

It is unlikely that there will be a single solution. The state may choose several of the possibilities listed and may identify others that are deemed workable within its system. There are other options, of course. Redesign the existing system. Throw it out and start over. What the state chooses to do will be determined only after careful consideration of the alternatives and an evaluation of the feasibility of each. The system ideals, mandates, purposes, standards, structures, state climate, and consequences will all factor into the decision the state will make.

State B

State B underwent significant educational reform in 1994 when the State superintendent marshaled stakeholders across the state to develop a set of high standards for which the educational system would be held accountable. Emphasis on standards, assessment, and accountability for all students became paramount. Standards were developed in Language Arts, Mathematics, Science, and Social Studies. Educators across the state were informed that these standards would be the basis for future rewards and sanctions. In 1997, a new testing contract was awarded for testing at grades 3, 5, 8, and 10. The tests included multiple choice items, short answers, and extended response items. Most of the short answer and extended response items were developed by teachers in the state. The testing program also includes the mathematics and reading sections of the SAT in order to provide comparisons to national norms. Data were to be reported at the student, classroom, school, and state levels.

In 1998 it was decided that students would have to exceed a cut score on the grade ten measure by 2001 to be eligible for graduation. The State released its initial inclusion guidelines and makes annual adjustments in this reference tool, most recently incorporating a scannable form to document the accommodations for each student. Accommodations up to a certain point were acceptable for inclusion of scores in totals and beyond that point the scores could not be aggregated. However, IEP teams were charged to document what each student needed without regard to whether the score would appear in the aggregate. Rationales for exclusion from testing had to be documented by IEP teams as early as 1995. "Head count audit teams" monitored the written rationales when exclusions exceeded 5% of students.

Early in 1997, the State recognized the need for aggressive work on an alternate assessment in order to meet the July 1, 2000 deadline. By the end of summer 1998, the state standards had been interpreted and "bridged" to encompass functional skills for students with more severe disabilities. A portfolio approach was selected and pilot-tested in the 1998-99 school year. The second pilot test year (1999-2000) was underway at the time of this writing.

The "gray areas" became more evident as the requirement to pass the state tests for graduation loomed and the IDEA ’97 requirements for inclusive testing became more apparent to teachers and parents. Advocates began to ask more specific questions about the appropriateness of the standards, accommodations, and the particular testing approaches for students with disabilities. They wondered whether the concept of "alternate assessment" should apply to these students. Specifically, parents and teachers dealing with students who are deaf began asking whether students should be demonstrating competency using American Sign Language (ASL) rather than using English, since ASL would be the primary language they would use for the rest of their lives. Those working with students who were blind questioned the appropriateness of some of the standards and test items, e.g., those having to do with maps and graphs. They also

questioned the requirement for "reading," since blind students were increasingly using scanners that translated written material into audio. However, scanners were only on the list of accommodations that led to non-aggregated scores.

Students with significant physical impairments presented another issue, since they often could not complete the test in the time allotted and, even if given extended time, would tire significantly. A proposal was made to provide a shortened version of the test and extrapolate scores. However the Technical Advisory Committee raised concerns about the extent to which the full range of content could be sampled, about the reliability of a shortened test and about the potential loss of credibility (i.e., perception that these students had only to meet lower standards). Without further research, the decision would have to be delayed.

Questions were raised about the extent to which tests could be accommodated or modified for more mildly involved students as well. For example, should the reading standards apply to a student with a learning disability in visual processing or would it be appropriate for such a student to demonstrate competency in extracting meaning from written material in another way (e.g., by having the reading material read to him or her). One of the questions was whether this would only be true if the disability was a life-long disability for which the accommodation would always be used.

In addition, legislators, local superintendents, and many parents were vocal in suggesting that off-level testing would provide more accurate estimates of a student's status for students whose reading levels were significantly lower than the grade level being tested. The state's technical advisory group recommended against using off-level tests since the tests were not equated across grade levels and therefore could not give an accurate picture of how the State was doing with its students at a particular grade level.

Finally, there were questions about which accommodations could be used for what sections, for example, a scribe was not allowed for the writing test prompt items but was acceptable for multiple choice items. Informal surveys of teachers revealed that the IEP teams had little understanding of the state standards or accommodation guidelines nor were they using consistent procedures for making decisions.

The emerging reality for this state is portrayed in Figure 2. This state is working hard to base its decisions on a balance of research and best available practice information. Therefore, the wide gray area relates to the extent to which stakeholders have confidence in the decisions they are making. Modified tests, tests that measure content irrelevant to the student's goals for the future, and measures that assess skills far beyond a student's capacity are considered gray areas.

Figure 2. Gaps between "Ideal" Assessment System and the Emerging Reality for State B

	An Emerging Reality in State B
	The general education test is appropriate, no accommodation needed and scores are comparable
	The general education test is appropriate with accommodations for which educators, measurement personnel, parents, and the public have confidence that the scores are comparable

	The general education test and the general content and performance standards are appropriate, but the test modifications are so drastic that they are perceived to change the content, and scores are not considered comparable
Gray Area	The general education test is appropriate with acceptable accommodations, but there is some discomfort about the long term relevance of the standards being tested for specific students
	The general test and standards are all that are available because the alternate assessment does not address the students' curriculum. However, regardless of test modifications, the scores for some students barely reach the random score level
	The alternate assessment is appropriate

This state is working on its gray area challenges. An advisory committee consisting of three large subcommittees dealing with deafness, vision impairments, and milder disabilities is looking at the accommodations guidelines to see whether they can be extended. Recommendations are expected before the Spring 2000 testing.

Accountability at the student level might be delayed because of public outcries. However, if the stakes remain, students who take modified tests or even off-level tests might be considered to have "passed." Finally, the state might take another look at not only the tests, but in light of the need for more inclusion, also at the state content and performance standards. The state standards might be revised and expanded to include not only the usual academic skills but also more compensatory academic skills for students who need life-long accommodations to function post school. Finally, additional access skills might be added to the standard curriculum and the state testing program for all students so that all students have an opportunity to exhibit more of what they know and can do in ways that are relevant to post-school life.

The Context of Gray Area Assessment Issues

National Reform Movement

We are in the midst of nationwide school reform. Schools in all 50 states are undergoing revolutionary change. Across the country, legislators, policymakers, educators, parents, and concerned citizens are working together to ensure that all children in our public schools develop the knowledge and skills necessary for them to lead productive and fulfilled lives in the 21st century. We, as a nation, are committed to a vision in which all students learn to high standards. To track our efforts, we are placing increased emphasis on measuring what students know, understand, and are able to do.

The vision is that our improved accountability systems will be used to measure progress and to plan improvements. In this model of school reform, we set content and performance standards, design curriculum and instruction to teach and learn to these standards, and then administer aligned large-scale assessments to measure our progress. Finally, we use data from these assessments and other data sources to adjust our efforts and ensure that all students in all schools succeed. We assume that this model will generate a cycle of continuous improvement.

We face success mixed with unanticipated challenges. As data-driven continuous improvement is implemented, we are learning more about what we expect and whom we include when we say *all* students. We are seeing unprecedented public interest in how our schools are doing. This interest has focused attention on how we measure, how we report, and how we use the data from large-scale assessments.

But a serious challenge has emerged. We are experiencing disharmony between high standards and all students. *All* students are expected to reach high standards and the accountability system is being used to identify areas of curriculum and instruction that the schools must improve. However, states and districts have identified areas where large-scale assessment systems seem lacking in their ability to assess and report what *some* students know and can do. These students include special populations such as students with disabilities, English language learners, and disadvantaged students.

As a nation we have committed to teaching every student who comes through the school door. We cannot change who these students are but we can improve the system that receives them, educates them, and assesses their performance. As we identify areas where large-scale assessments seem lacking for some students, we must address the problem.

National Reform Context for Students with Disabilities

The amended Individuals with Disabilities Education Act 1997 (IDEA '97) and other federal legislation (specifically Title I of the Improving America's School Act [IASA]) call for assessment and reporting that accounts for every student. Title I assessment requirements are to be used as the primary basis for school and district accountability, and must include all students, with appropriate measures for students with disabilities and English language learners. Assessments must be aligned to content and performance standards, and provide data for school profiles, including disaggregated data.

IDEA '97 requires States to establish goals for the performance of children with disabilities that are consistent, to the maximum extent appropriate, with other goals and standards for children established by the State. It also requires that children with disabilities are included in general State and district-wide assessment programs, with appropriate accommodations, where necessary, or alternate assessments for those children who cannot participate in State and district-wide assessment programs. When States report assessment results to the public, they must include aggregated data that include the performance of children with disabilities together with all other children; and disaggregated data on the performance of children with disabilities.

The issues surrounding alternate assessment have been articulated elsewhere (Olsen, 1998; Ysseldyke, Olsen, & Thurlow, 1997), and states are working hard to meet the deadline for development and installation of alternate assessment systems (Thompson, Erickson, Thurlow, Ysseldyke, & Callender, 1999; Warlick & Olsen, 1999). In addition, nearly all states have revised their guidelines for IEP decision making and their accommodation guidelines (Thurlow, House, Boys, Scott, & Ysseldyke, 2000).

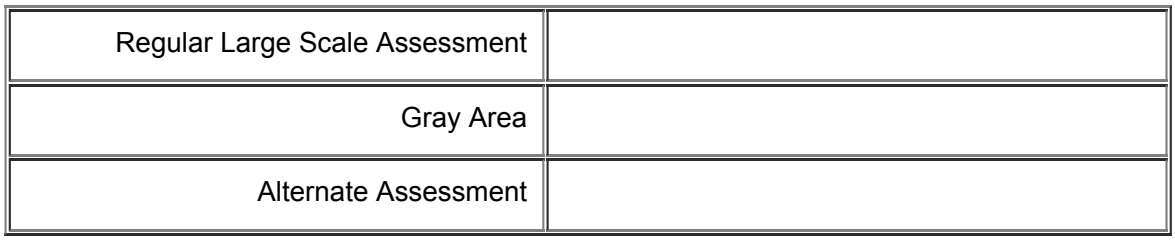
This paper focuses specifically on an area between the regular assessment and the alternate assessment. This is an area where an alternate assessment is inappropriate, but the large-scale assessment does not seem to work for the student with disabilities even with accommodations. We include as gray the area where the large-scale assessment has been modified to the point

where the results cannot be included comparably in state summary data.

A Model of the Gray Areas of Assessment Systems

In early work on accommodations and alternate assessment, the gray area of assessment was conceived as a well-defined area between the general education large-scale assessment and the newly mandated alternate assessment (see Figure 3). Many states had an assessment system already in place, were expanding this system to address state adopted content and performance standards, and were identifying accommodations to address these limited gray area concerns. It logically appeared that the alternate assessment would be for students who could not take the general education assessment under any conditions, and all other students would be included in the large-scale assessment.

Figure 3. Gray Area as Well-Defined Area between General Education Large-Scale Assessment and the Alternate Assessment



However, as the states have proceeded in development, they find there are areas where large-scale assessment systems, even with accommodations and modifications and with the development of alternate assessment options, are inadequate for showing what all students know and can do. The gray areas of assessment appear to be more complex and challenging than first conceived.

Each state seems to encounter a unique version of this problem. This was evident in the two state examples presented earlier. Across the states, the gray areas of assessment systems are affected by how states differ on key components of the accountability system. The beliefs and assumptions that shape these components uniquely affect the gray areas. Many systems are driven by state legislated mandates that reflect differing state contexts and demography, varying philosophies on equity and excellence, and different assumptions about the purpose of the assessment system. Similarly, content and performance standards in each state reflect differing values and understandings, and affect gray areas differently from state to state. The assessment process and format, both the allowable accommodations and the testing program formats, affect gray areas, and again, vary based on assumptions and purpose. Finally, the determination of high and low stakes for students, schools, and systems influences which students are affected by gray areas.

More explicitly stated, the influencing factors include ideals, mandates, purposes, standards, structures, state climate, and consequences. The ***ideals*** are expressed in the underlying beliefs and assumptions of the system as well as through community based approaches to equity and excellence. ***Mandates*** compel the system to meet broad values from federal legislation, state

statutes, rules and regulations, the leadership of the superintendent or commissioner, and state board of education decisions. Stated ***purposes*** for large-scale assessments, while unique to each state, provide a framework for the system and range from education accountability and student progress monitoring to the planning and improvement of schooling. Similarly, state content and performance ***standards*** established within states reflect varying values and understandings and affect gray areas differently from state to state. These ideals, mandates, and purposes interact with assessment system ***structures*** such as: grades, subjects, and levels tested; conditions of administration (accommodations, modifications, frequency, etc.); criterion vs. norm referenced or standards based designs; methods of scoring and reporting; and type of test such as multiple choice, extended response, or performance assessment. Structures also refer to testing formats or modes of responding such as paper pencil, computer based or assisted, oral responding, and so on. These components not only work within a ***climate***, that is, the state's diversity, economics, geography, demographics, and politics, but it appears that stakes or ***consequences*** to students, schools, and personnel further compound the effects such characteristics have on the shape and size of the system's gray areas.

The gray areas tend to change over time. As states proceed with school reform and the continuous improvement cycle, their understanding of what they want to accomplish, and how to get there becomes more refined. As this occurs, the gray areas change from year to year.

Questions to Define and Address Gray Area Concerns

Based on varying components of accountability systems, we have posed five interrelated questions that can help us define and address gray area concerns in any state or district at any point in time. We explore each of these questions by first identifying the context of each question, and then identifying issues to address and discuss. The five questions are:

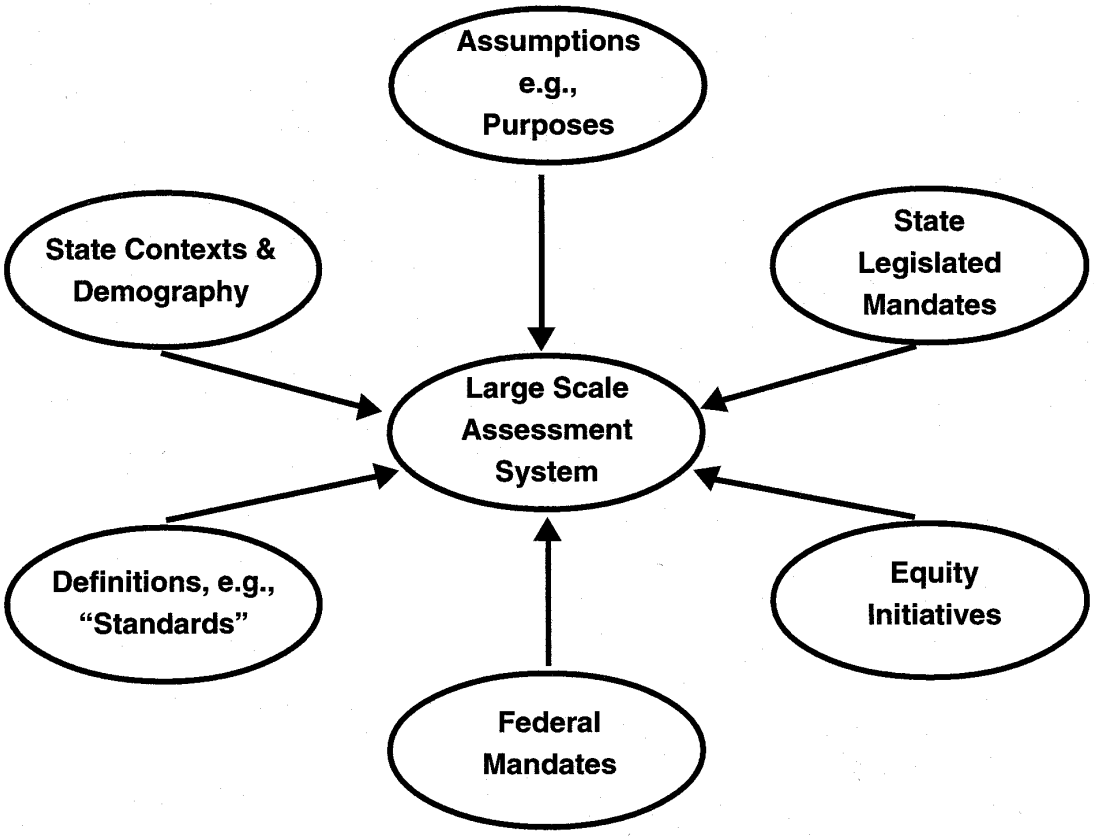
1. What is driving our large-scale assessment programs, and how does it affect gray area concerns?
2. How does a state or district approach to content and performance standards affect gray area concerns?
3. How do test accommodation and modification policies affect gray area concerns?
4. To what extent do assessment formats affect gray area concerns?
5. How does the nature of the high and low stakes accountability system affect gray area concerns?

What is Driving Large-Scale Assessment Programs, and How Does That Affect Gray Area Concerns?

Context of the question. The forces that shape large-scale assessment programs come from multiple sources (see Figure 4). They include federal, state, and local mandates that reflect beliefs and values at each level. These beliefs and values may result in a relative emphasis on equity and excellence, or they may play out in technical approaches to measurement favored by

policymakers or assessment experts. In some states, there are legislated mandates to use national norm-referenced tests as a measure of school accountability, sometimes customized to align to state content standards. In other states, mandates require new criterion-referenced tests aligned to state standards, or development of a performance assessment system to measure progress toward standards. In some states, these tests are used for a variety of purposes (e.g., individual student assessment, instructional planning, school improvement, and systems accountability). These choices reflect differences in state contexts and demography and result in very different implications for the gray areas of assessment.

Figure 4. The Forces that Shape Large-Scale Assessment Programs



Issues to consider and discuss. The first set of issues we need to address is the underlying assumptions and definitions that serve as a basis for a state or district program. To what extent is the purpose of a particular assessment compatible with a mandate for a fully inclusive system? Do we think the students are a problem because they do not fit the system or do we think our measures are the problem because they do not fit the students?

What is driving most of our decisions: state standards, or test norms and standardization? If standards, are we focused on assessing the current state of affairs and what students *have* been learning? Or do we clearly focus on what students *should* be learning, including students who traditionally have not been taught within the general education curriculum? Is this confusion just a timing and developmental issue in the reform process? Will it go away as we change state curricular approaches and begin to instruct all students on state standards?

Are our guidelines for participation in the alternate assessment based on the characteristics of our general test or are the guidelines based on student characteristics and needs? If the former,

what are the implications if we change our test? What are our assumptions about such things as limits on student potential, expected uses of data, political factors, and costs?

How have we addressed the tendency for parents and special educators to "protect" students from what are perceived as the personal risks of inclusive assessment? Are we aware of any real dangers of unintended and negative outcomes of inclusive assessment, and if so, how have we addressed these dangers?

How are these issues affecting our thinking about the alignment between our tests and our students? Do our challenges in matching test and student relate to particular groups of students? And if so, are there particular challenges with: all students with disabilities? Students with mild disabilities? English language learners? Poor readers? Students who have not had opportunities to learn? Students who have never learned how to take a test or who have test phobia?

How Does a State or District Approach to Content and Performance Standards Affect Gray Area Concerns?

Context of the question. In general, content standards identify what students should know and be able to do. Performance standards typically define the level of performance expected on the content standards, often with an absolute score on some type of assessment, sometimes called a "cut score," that defines whether a student demonstrates content knowledge and skills to the level required by the state.

States vary in the approach and the degree of flexibility built into their standards. Some states have committed to an approach to school reform that emphasizes basic skills in math and reading, or other core content, and they have an assessment system designed to measure highly specific content standards to highly specific performance levels in prescribed settings. Other states have developed a cross-disciplinary approach to standards, emphasizing demonstration of skills and knowledge in a variety of disciplines, with flexible performance settings and levels. Still other states fall somewhere between, with highly specific content standards across multiple disciplines.

Some researchers on national standards implementation suggest that performance standards should identify the environments in which knowledge and skill should be demonstrated, defining specific use of that knowledge, as well as defining the expected quality of performance (Marzano & Kendall, 1997). Whether and how a state defines performance levels in multiple settings profoundly influences the gray areas.

How the states have addressed extending or expanding the state content and performance standards for alternate assessment populations also affects gray areas. Some states have developed separate content and performance levels for these populations; other states have defined core competencies within their state content standards toward which all students work.

Issues to consider and discuss. How do these content and performance standards affect the gray area? Do we hold students accountable for standards in those content areas considered "basic" or do we hold students accountable for content knowledge and performance levels

across a wide range of content? What have our stakeholders defined as "basic?" How flexible are those definitions for students with a full range of unique needs? Do these definitions contribute to students being left out of standards-based instruction and assessment? Do the performance standards create an artificial gray zone due to narrowly defined context prescriptions? What about standards that make no sense for students with certain characteristics, e.g., specific listening skills for a deaf student or specific observation skills for blind students?

Are our content standards focused on traditional subject areas such as mathematics, science, history, geography, language arts, fine arts, and foreign languages? Do we have separate standards for general reasoning skills, including decision-making and problem-solving? Or are those skills embedded in our core content standards? Do we have separate standards for work related skills such as time management, teamwork, or resource management? Have those skills been embedded in core content and performance standards? How do these variations affect how students with unique needs "show what they know?"

Is our assessment designed to assess the surface nature of the standard, or the depth of the concept behind the standard, e.g., school-based knowledge only or the life role in which that knowledge would be applied? For example, must we assume that a content standard about a math operation really means that a student must perform that operation in his or her head or that the student is expected to be able to have a way to get an answer, perhaps using an adaptive device? Are we assessing the concept and need behind the content standard or the literal phrasing of the standard?

Do our performance standards allow demonstration of knowledge and skills in a variety of settings? Have we defined levels of performance that can differentiate where students are in their progress toward achieving standards, in whole or in part? Do we have options to allow a variety of assessment techniques as part of regular classroom standards-based instruction? How do those options relate to large-scale assessment of student knowledge and skills?

For example, could a student who is working on communications or mathematics content standards in a transition/work based setting demonstrate mastery of these content standards through a performance assessment in the work place? Can the same work place assessment measure student progress toward standards in core academics, thinking and reasoning, and work related skills? Does that differ for student vs. system accountability? Could results be aggregated? Under what conditions would this be considered?

How Do Test Accommodation and Modification Policies Affect Gray Area Concerns?

Context of the question. There are several accommodation and modification issues, many of which overlap with the standards issues above. There are wide variations in state policies and guidelines for assessment, with variation across states in what are considered to be "standard" accommodations, even for the same nationally standardized test (Thurlow, House, Boys, Scott, & Ysseldyke, 2000). The research base is not clear on these distinctions, and although most states make these decisions in conjunction with test publishers, the decisions tend to be based on opinion, rather than solid research (Tindal, 1998). These decisions clearly affect which students are affected by the gray areas.

One model commissioned by the State Collaborative on Assessment and Student Standards,

Assessing Special Education Students Study Group III (SCASS ASES) makes a distinction between accommodations and modifications along a continuum (Tindal, 1998). In this continuum an accommodation was defined as a "change that (a) provides unique and differential access (to performance) so certain students may complete the tests and tasks without other confounding influences, but (b) does not change the nature of the construct being tested." Such changes typically are designed for specific individuals and for particular purposes. The concern is the extent to which the basic construct has been changed by the accommodation. The SCASS has defined an accommodation as a change that does not affect the construct, and results in a score that can be aggregated, and a modification as a change that does change the construct being measured, with limited ability to be aggregated or included in summary statistics.

Issues to consider and discuss. What policies and procedures are in place to ensure that accommodations that level the playing field and do not change the performance standard are available, as appropriate? Is the issue whether the accommodation allows the student to show what he or she knows and is able to do? Are there disability-specific issues?

Can we substitute tasks that assess the same skill or concept and still treat the data as part of the whole? This might be especially important for students with life-long disabilities for which they will have to compensate their entire life. For example, in a map-reading task, can a blind student be asked to use whatever techniques she will use the rest of her life to locate a place, determine a distance, or do a comparable skill to whatever is being measured? Can a student with a permanent decoding disability demonstrate how he will get information from a typed paper by using a scanner?

If the stakes for students are low (e.g., not related to promotion or graduation) are modified tests acceptable even if they change the constructs being measured? Can we use the results from modified tests as part of an accountability measure even though we do not feel we can include scores in a report of aggregated state averages? What are the intended and unintended effects of assigning "0" scores to modified tests used in an accountability measure? What other reporting issues are affected by test modifications?

With the increasing availability of research on the effects of accommodations, will we see what are now defined in practice as modifications becoming accommodations or will we find less justification for accommodations? Do we need to rethink the issue of accommodations and modifications for more students—students without IEP, 504, or ELL documentation? What is the impact of expanding this consideration?

To What Extent Do Assessment Formats Affect Gray Area Concerns?

Context of the question. Testing programs differ across the states and districts. Some states rely on a single large-scale assessment for accountability. Other states use multiple approaches, with some measures used statewide, and others developed locally. States use norm-referenced, standards-based, criterion-referenced, multiple choice, short answer, extended response, responses to prompts, on demand performance assessments, and portfolio assessments, off-the-shelf, customized, state developed, and teacher developed assessments. They use them alone or in combination, in a variety of settings, for high stakes or not. Some states require participation in the state test, while others allow local options (Olson, Bond, & Andrews, 1999).

Each of these variations may affect gray areas.

Issues to consider and discuss. Is it possible that there are characteristics of testing programs that complicate the gray areas? Do we have gray areas only when we have some evidence that students know something but our measures cannot show it, that is, a student does not meet a standard only because of the way it is measured? Is this a validity issue that would require us to consider and account for a student's disability as a source of error variance in our testing program? Should we concede the possibility that an assessment program can never be truly inclusive? Have we considered multiple options for demonstrating achievement for all students? Is there an interaction between the type of test and the extent to which tests are inclusive? Are multiple choice tests more or less of a problem than short answer or extended response items? What about writing prompts? Are time limits increasing the problem? Would portfolios help or hinder our inclusion problem? If we are trying to use a norm-referenced measure, do we have a greater problem than when we are standards-based? Can we use a norm-referenced test when: (a) students with disabilities were not included in the norm sample, or (b) certain accommodations were not provided?

Do we have options to allow a variety of assessment techniques as part of classroom instruction? If so, how have we provided for validity and reliability? And what are the benefits and risks of teacher assessment related to issues of teacher low expectations and misguided protection of students with disabilities?

How Does the Nature of the High and Low Stakes Accountability System Affect Gray Area Concerns?

Context of the question. The state system of assessment and accountability may include "high stakes" for the system, school, or the individual student. These may include rewards or sanctions for school improvement at the systems and school level; or promotion or graduation stakes for individual students. How these stakes have been defined and implemented also affects the definition and impact of gray areas.

Some states have determined that if the test is used only for system accountability, most students should attempt the test, even if their scores would be at the minimal or chance level (Thurlow & Thompson, 2000). If all schools include all students in the assessment system, then the relative scores of students previously excluded should apply equally across schools, and have limited effect on accountability indices. Not all states have adopted this policy, and this solution to "gray areas" for systems accountability continues to be debated.

But when high stakes exist for the individual student, the gray areas cause profound problems. Diploma options and other graduation policies are controversial topics embroiled in concerns about the meaning of a high school diploma and the potential long-term effects of not receiving a diploma. The consequences of graduation and diploma policies last well beyond the time of high school attendance. Yet, these concerns must be weighed against the desire to have a high school diploma mean something—that a student has mastered specific knowledge and gained specific skills. Balancing these against a desire to be fair to students and to not harm them create significant challenges for states today (Thurlow & Thompson, 2000).

Low expectations for students with special needs have created some gaps in knowledge and skills for students currently in our public schools. The short and long term problems of opportunity to learn for "all students" are linked to assumptions about who "all students" are, how standards apply to "all students," and how "all students" can demonstrate what they know and can do. Like our other questions, the discussion of high and low stakes is interconnected to what drives the assessment system, what the state's content and performance standards are, technical and format issues related to the assessments themselves, as well as how the stakes have been defined.

Issues to consider and discuss. The issue of stakes seems to exacerbate the gray areas. Are there any conditions under which the gray areas can be ignored? For example, if the issue is school level accountability can we simply say that all students count and if they cannot take the test, they count as a zero (or whatever the lowest level is)? Can we simply ignore students who really cannot take the test because the total number of such students is small and would not affect our averages? Or are there other reasons to include them? Can we just make adjustments in our accountability measures and forget the actual assessment process? What about for accountability measures used for school and program improvements? How do the issues change if we are trying to assign a level to an individual student?

Does the high stakes purpose for which the testing program exists affect gray areas? For example, is our flexibility greater or more restricted under high stakes for schools versus high stakes for students? Is this true when the only decisions relate to instructional planning at a student or system level, but there are no specific consequences?

Do we have multiple methods for student demonstration of progress toward achieving standards or do we have one high stakes assessment? What about re-takes, re-scores, appeals and alternate evidence such as juries and portfolios based on the same standards? How can we address the gray area without actually lowering standards or even appearing to do so?

What policies and procedures are in place to align the IEP goal-setting process to content and performance standards, to assessment of those standards, and to high stakes? Assuming that alignment, can the IEP replace the assessment program in ways that meet high stakes requirements, but still provide information for aggregation? How can that be developed to ensure high expectations for all students?

Are the gray areas related more to opportunity to learn, seat time, Carnegie units, and other issues rather than to assessment? How do we develop and implement tests that are appropriate for all without "watering down" the high content standards and thinning the rigor of the performance standards? Are we measuring progress toward high standards and using results to identify what should be taught, or are we content with measuring lower expectations? For example, if they cannot read, do we test reading at their reading level or give them a test they cannot complete to demonstrate where they really stand? How does that answer change when the purpose of the test is for system accountability as opposed to student accountability?

How are these issues related to the increasing number of states installing high stakes tests for students, such as needing to pass a test to obtain a diploma (Thurlow & Thompson, 2000)? What does the diploma mean? Do our transcripts reflect actual student progress toward standards, or do they simply reflect a "met/not met" criterion? Should we be focusing on credentialing as we consider transcripts vs. diplomas? Would credentialing benefit children with

disabilities by identifying accommodations, supports, and areas of need as well as areas of strength for post-secondary or work environments? What are the political and legal issues related to this approach, and could it be applied to all students?

Have we adequately prepared our communities, schools, teachers, and students for strict accountability on high standards for all students? Or are we charging the cost of school reform to the children caught in the gray areas? How do we make the transition to requiring success for all students while protecting students in systems where no guarantees of opportunity to learn were given? Should we begin by holding the system accountable first, and once that is in place, the students?

Conclusion

Up to now, most individuals who have been dealing with the problem of students who do not fit into an assessment system have assumed that the problem was with the students themselves—they were the gray area students. Some people went so far as to suggest that different assessments should be developed for these students—even though they realized that the students were working on the same general standards as other students and that the alternate assessment was inappropriate for them. Almost always, it was concluded that these students could not be counted in the accountability systems in the same way that other students were. This meant that systems did not count them or account for them.

As discussion progressed, more and more people realized that the problem really did not rest with the students, but rather was a function of the "gray areas of assessment." By reframing the concern this way, it is now possible for a district or state to consider its own context in addressing the issues that accompany the gray areas of assessment. The questions that can help states to clarify the issues for themselves focus on:

> **The assumptions and other factors underlying the large scale assessment programs**. What is driving large-scale assessment programs, and how does that affect gray area concerns?
>
> **The nature of standards**. How does a state or district approach to content and performance standards affect gray area concerns?
>
> **Participation and accommodations policies**. How do test accommodations and modification policies affect gray area concerns?
>
> **Assessment formats**. To what extent do assessment formats affect gray area concerns?
>
> **The stakes attached to the accountability system**. How does the nature of the high and low stakes accountability system affect gray area concerns?

As states begin to address these issues, it will become clear that the gray areas are not the same everywhere. The number of issues and nature of those issues are related to the state or district context, and will therefore be different in different places. Only by beginning this identification of relevant issues and responding to them can states and districts hope to avoid the criticism that their assessment systems do not account for every student within their public education system.

References

Marzano, R. J. & Kendall, J. S. (1997) *The fall and rise of standards-based education*. Mid-Continent Regional Educational Laboratory.

National Center on Educational Outcomes. (1999). *Forum on alternate assessment and "gray area" assessment*. Minneapolis: University of Minnesota.

National Commission of Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

Olson, J. F., Bond, L., & Andrews, C. (1999). *Data from the annual survey: State student assessment programs*. Washington, DC: Council of Chief State School Officers.

Olsen, K. (May, 1998). *Alternate assessment issues and practices*. Lexington, KY: University of Kentucky, Mid-South Regional Resource Center.

Thompson, S., Erickson, R., Thurlow, M., Ysseldyke, J., & Callender, S. (1999) *Status of the states in the development of alternate assessments* (Synthesis Report 31). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., House, A., Boys, C., Scott, D., & Ysseldyke, J. (2000). *State participation and accommodations policies for students with disabilities: 1999 update* (Synthesis Report 33). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. & Thompson, S. (2000). *Diploma options and graduation policies for students with disabilities* (Policy Directions 10*).* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Tindal, G. (1998). *Models for understanding task comparability in accommodated testing*. Eugene, OR: Behavioral Research and Teaching

Warlick, K. & Olsen, K. (1999). *How to conduct alternate assessments: Practices in nine states*. Lexington, KY: University of Kentucky, Mid-South Regional Resource Center.

Ysseldyke, J., Olsen, K., & Thurlow, M. (1997). *Issues and considerations in alternate assessments* (Synthesis Report 27)*.* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

NCEO is supported primarily through a Cooperative Agreement (#H326G050007, #H326G110002) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. Additional support for targeted projects, including those on LEP students, is provided by other federal and state agencies. Opinions expressed in this Web site do not necessarily reflect those of the U.S. Department of Education or Offices within it.